Evaluating Cloud Computing for HPC Applications

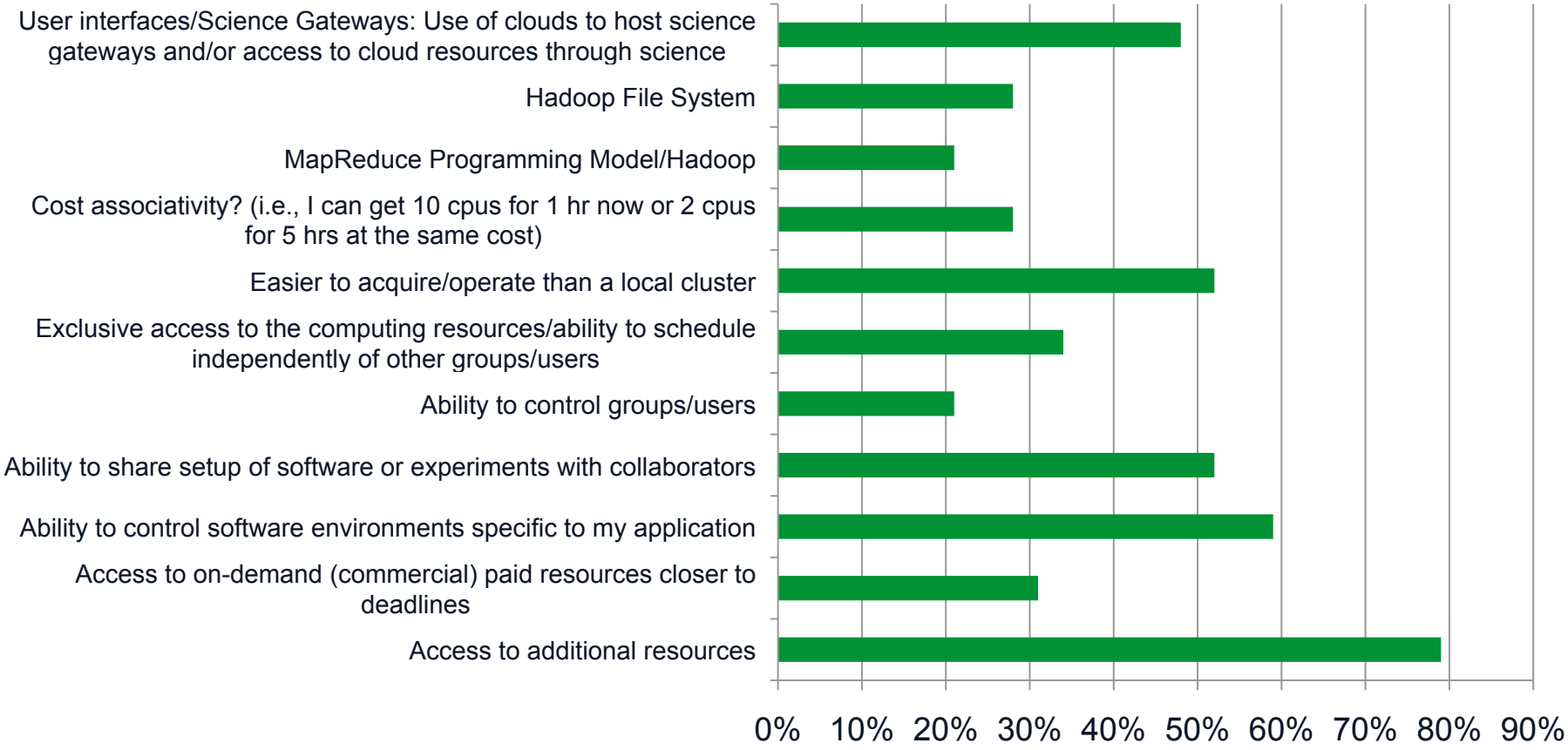
Lavanya Ramakrishnan

CRD & NERSC

Magellan Research Agenda

- **What are the unique needs and features of a science cloud?**
  - NERSC Magellan User Survey
- **What applications can efficiently run on a cloud?**
  - Benchmarking cloud technologies (Hadoop, Eucalyptus) and platforms (Amazon EC2, Azure)
- **Are cloud computing Programming Models such as Hadoop effective for scientific applications?**
  - Experimentation with early applications
- **Can scientific applications use a data-as-a-service or software-as-a-service model?**
- **What are the security implications of user-controlled cloud images?**
- **Is it practical to deploy a single logical cloud across multiple DOE sites?**
- **What is the cost and energy efficiency of clouds?**

Magellan User Survey



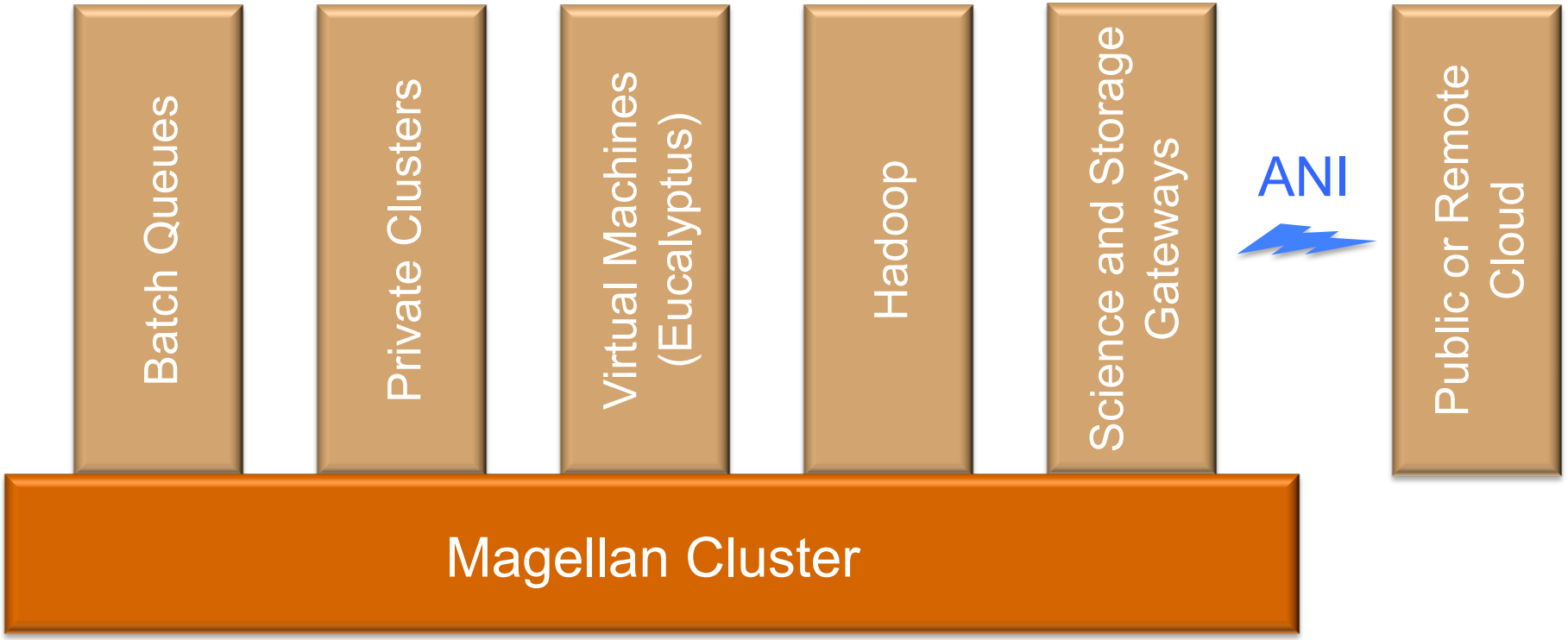
Program Office	
Advanced Scientific Computing Research	17%
Biological and Environmental Research	9%
Basic Energy Sciences -Chemical Sciences	10%
Fusion Energy Sciences	10%

Program Office	
High Energy Physics	20%
Nuclear Physics	13%
Advanced Networking Initiative (ANI) Project	3%
Other	14%

Cloud Computing Services

- **Infrastructure as a Service (IaaS)**
  - Provide unlimited access to data storage and compute cycles
  - e.g., Amazon EC2, Eucalyptus
- **Platform as a Service (PaaS)**
  - Delivery of a computing platform/software stack
  - Container/images for specific user groups
  - e.g., Hadoop, Azure
- **Software as a Service**
  - Specific function provided for use across multiple user groups (i.e. Science Gateways)

Magellan Software



Amazon Web Services

- **Web-service API to IaaS offering**
- **Uses Xen paravirtualization**
  - cluster compute instance type uses hardware assisted virtualization
- **Non-persistent local disk in VM**
- **Simple Storage Service (S3)**
  - scalable persistent object store
- **Elastic Block Storage (EBS)**
  - persistent, block level storage

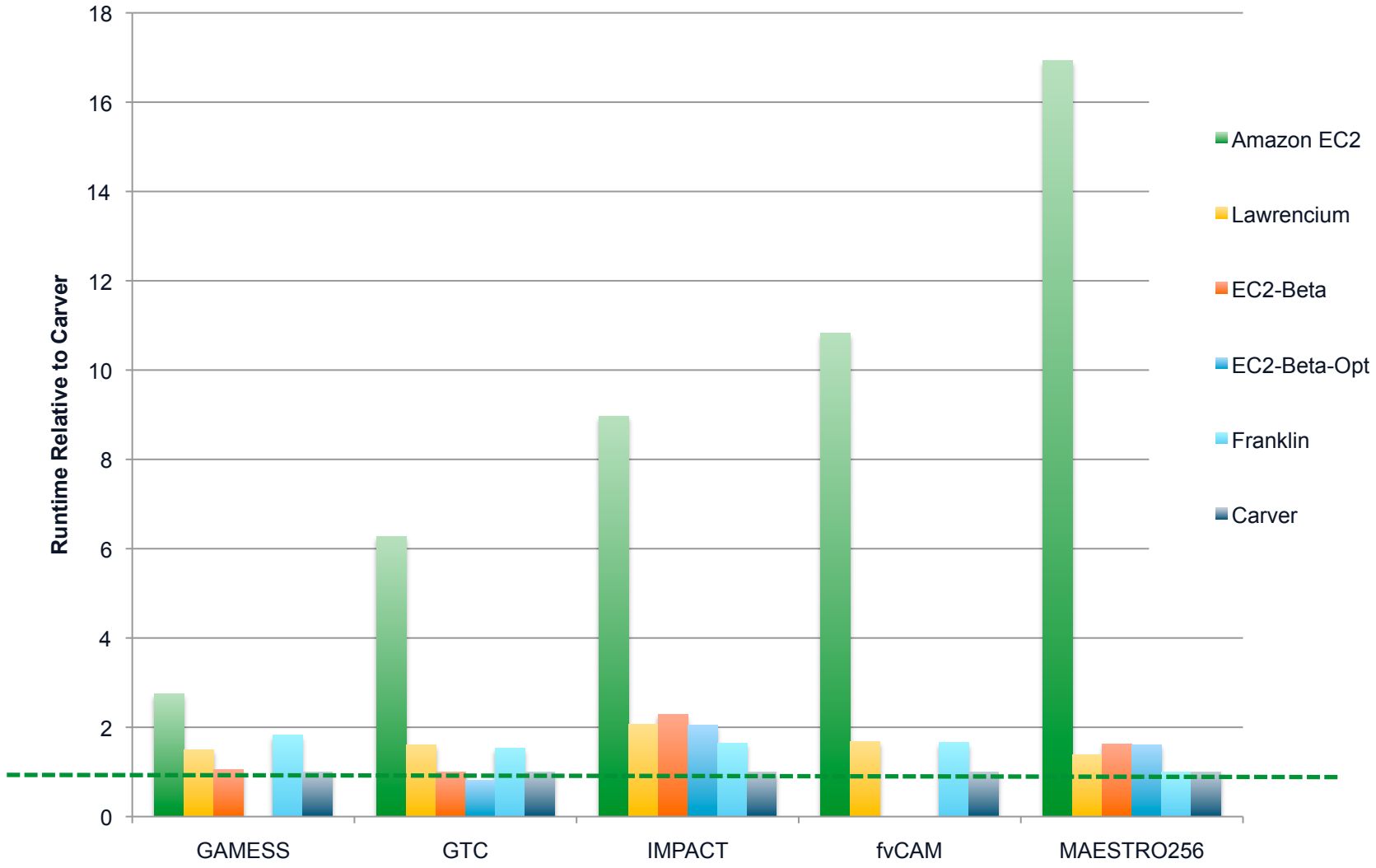
Eucalyptus

- **Open source IaaS implementation**
  - API compatible with Amazon AWS
  - manage virtual machines
- **Walrus & Block Storage**
  - interface compatible to S3 & EBS
- **Available to users on Magellan testbed**
- **Private virtual clusters**
  - scripts to manage dynamic virtual clusters
  - NFS/Torque etc

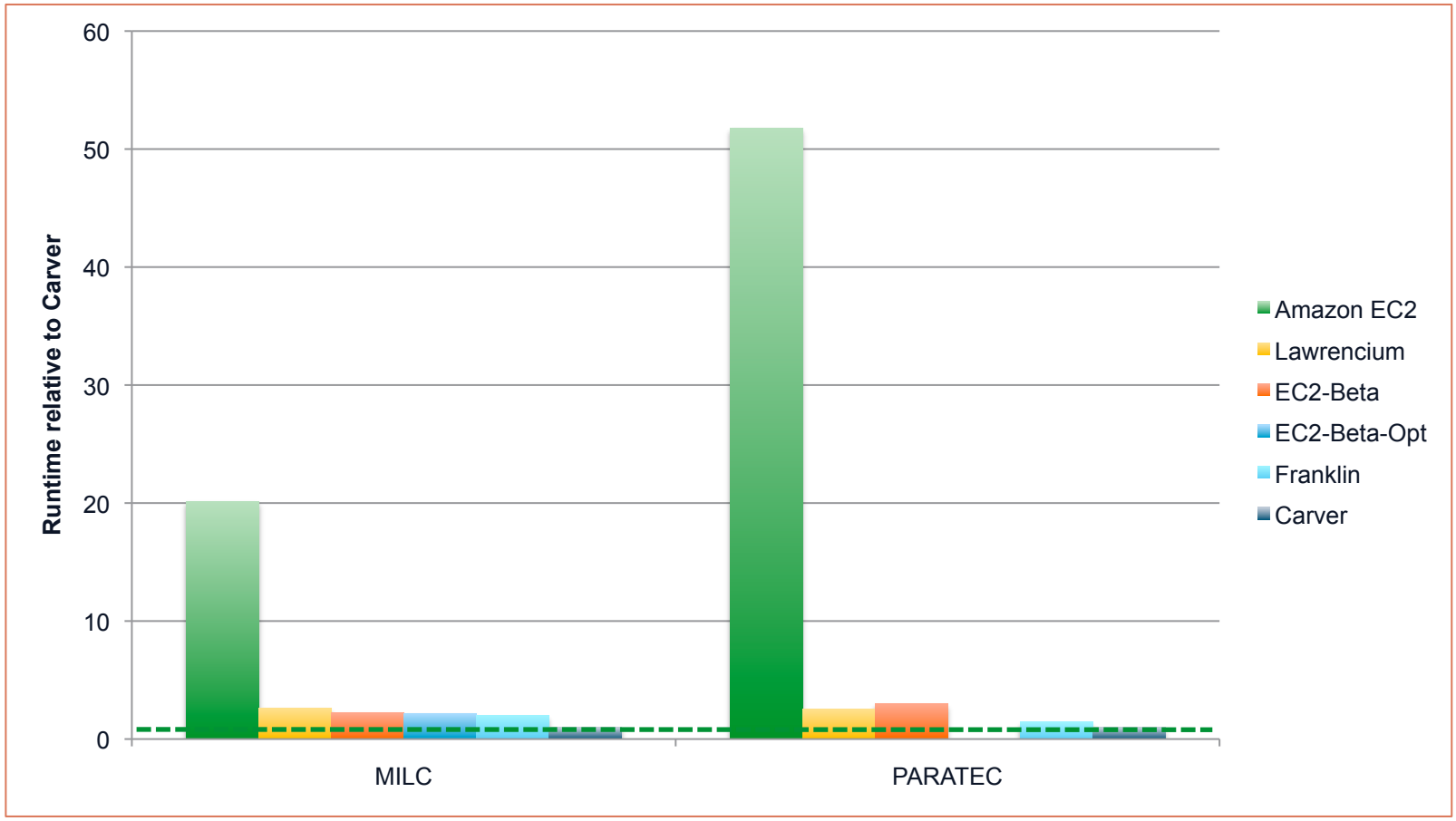
Virtualization Impact

- **Platforms**
  - Amazon, Azure, Lawrencium (IT cluster)
  - Magellan
    - IB, TCP over IB, TCP over Ethernet, VM
- **Workloads**
  - HPCC
  - NERSC6 Benchmarks
  - Applications Pipelines
    - JGI Supernova Factory
- **Metrics**
  - Performance, Cost, Reliability, Programmability

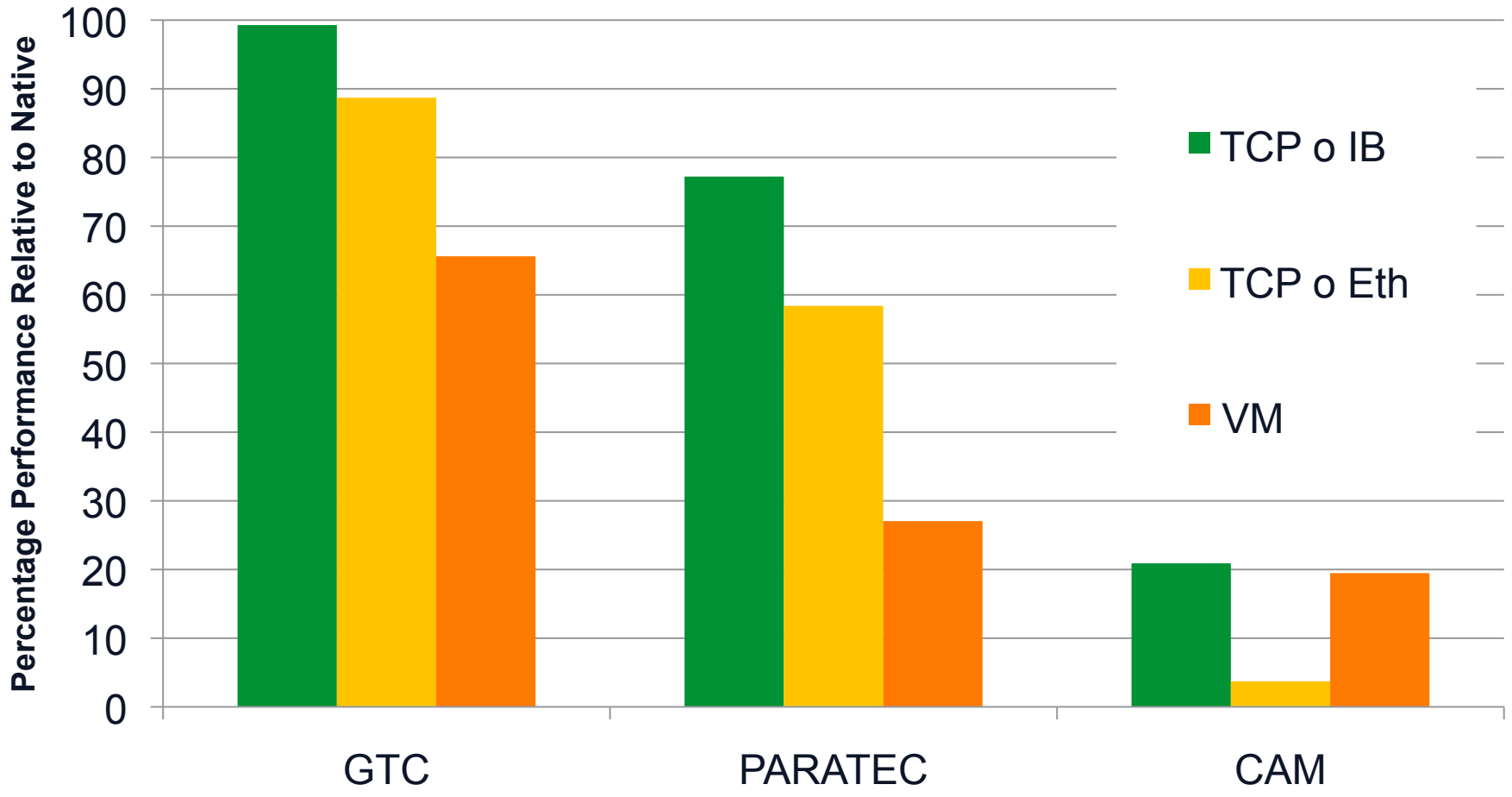
NERSC-6 Benchmark Performance [1/2]



NERSC-6 Benchmark Performance [2/2]



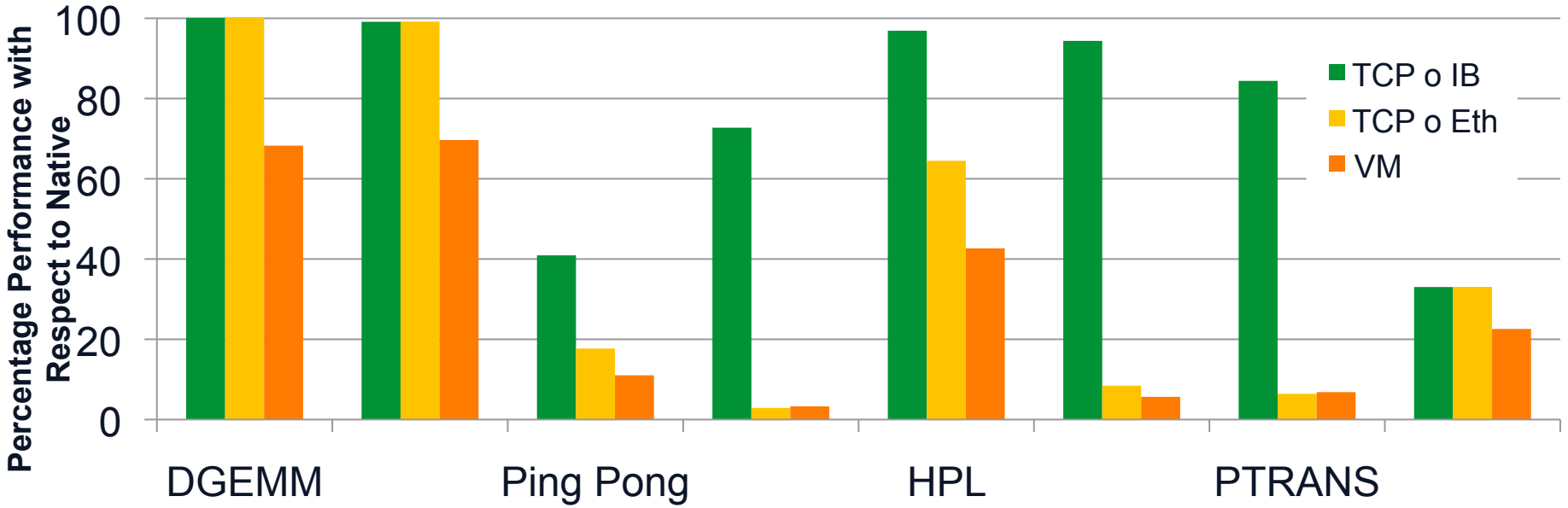
Magellan: NERSC6 Application Benchmarks



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB Lawrence Berkeley National Laboratory

Magellan: HPCC



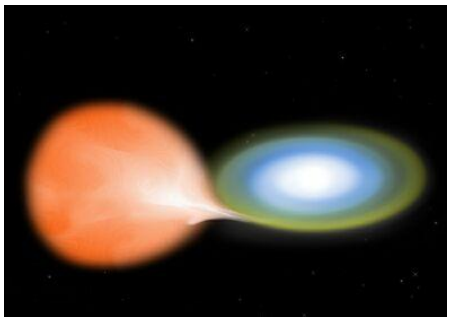
Performance-Cost Tradeoffs

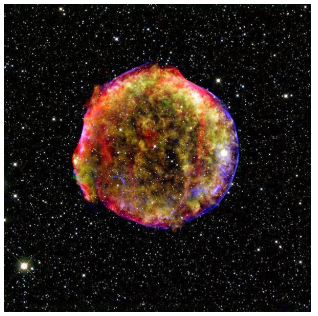
- **James Hamilton's cost model**
  - http://perspectives.mvdirona.com/2010/09/18/OverallDataCenterCosts.aspx
  - expanded for HPC environments
- **Quantify difference in cost between IB and Ethernet**

Application Class	Performance Increase/Cost Increase
Tightly Coupled with IO (e.g., CAM)	4.36
Tightly Coupled with minimal IO (e.g., PARATEC, GTC)	1.2

Nearby Supernova Factory

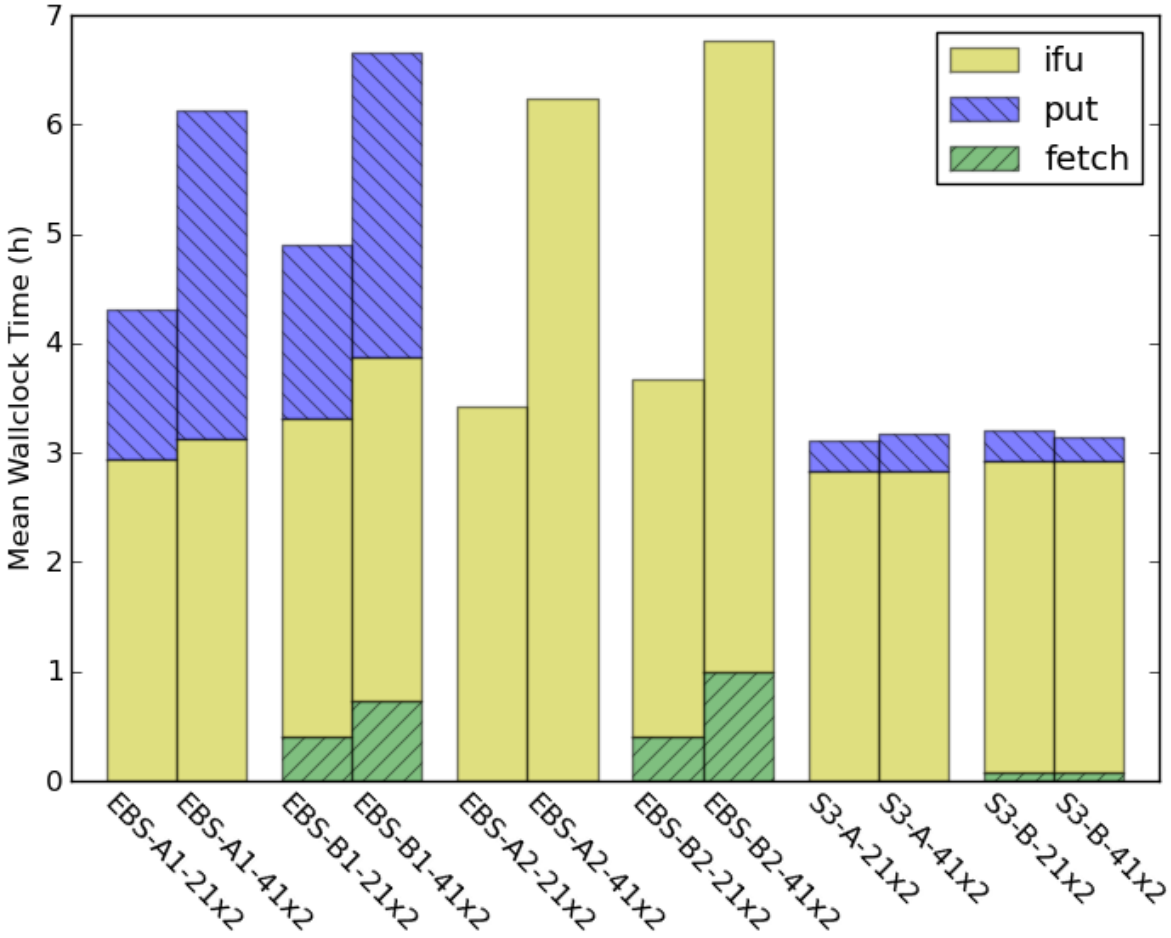
- **Tools to measure the expansion history of the Universe and explore the nature of Dark Energy**
  - Largest data volume supernova search
- **Data Pipeline**
  - Custom data analysis codes
    - Coordinated by Python scripts
  - Run on a standard Linux batch queue cluster
- **Cloud provides**
  - Control over OS versions
  - Root access and shared “group” account
  - Immunity to externally enforced OS or architecture changes



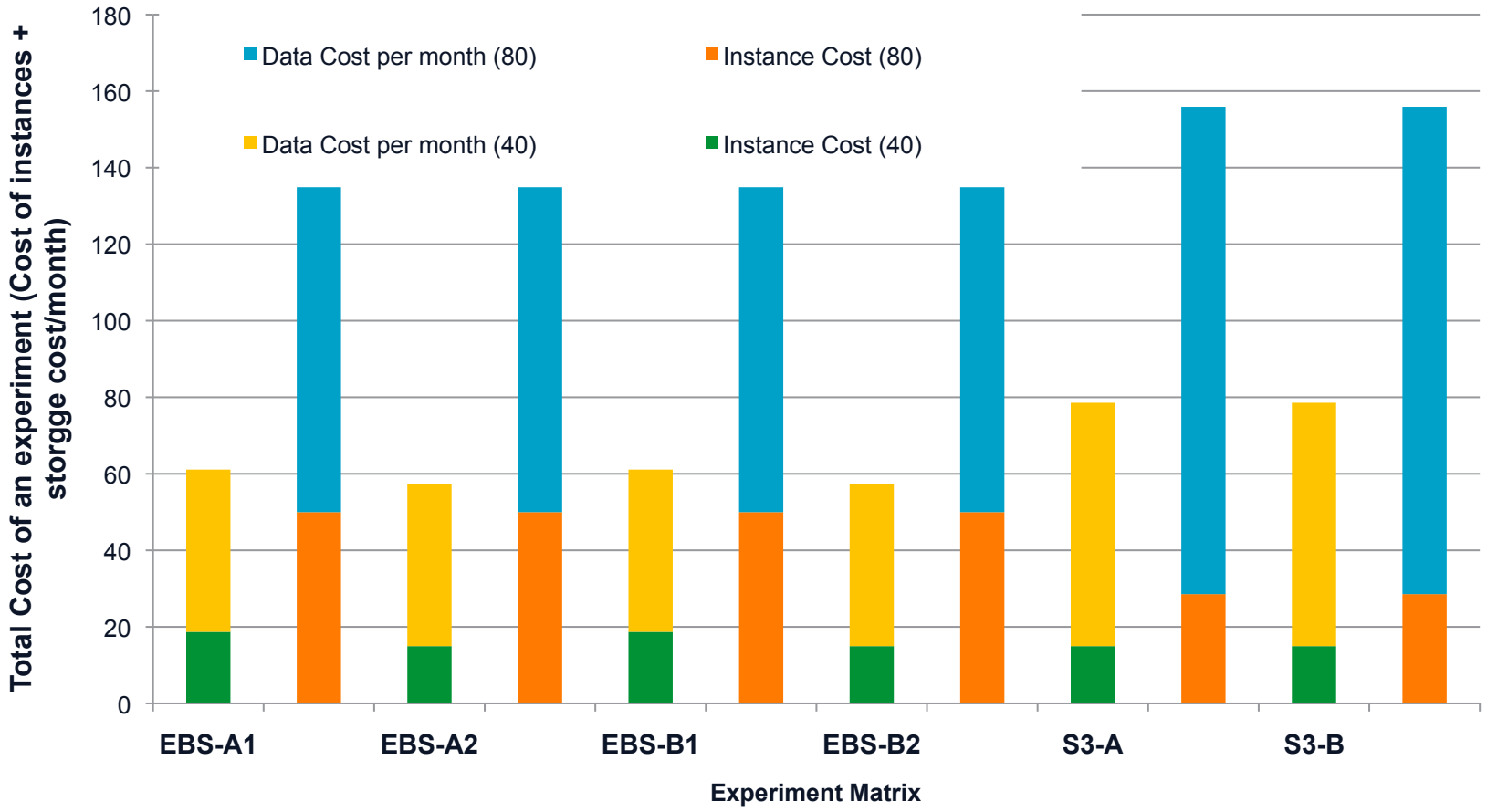


Experiments on Amazon EC2

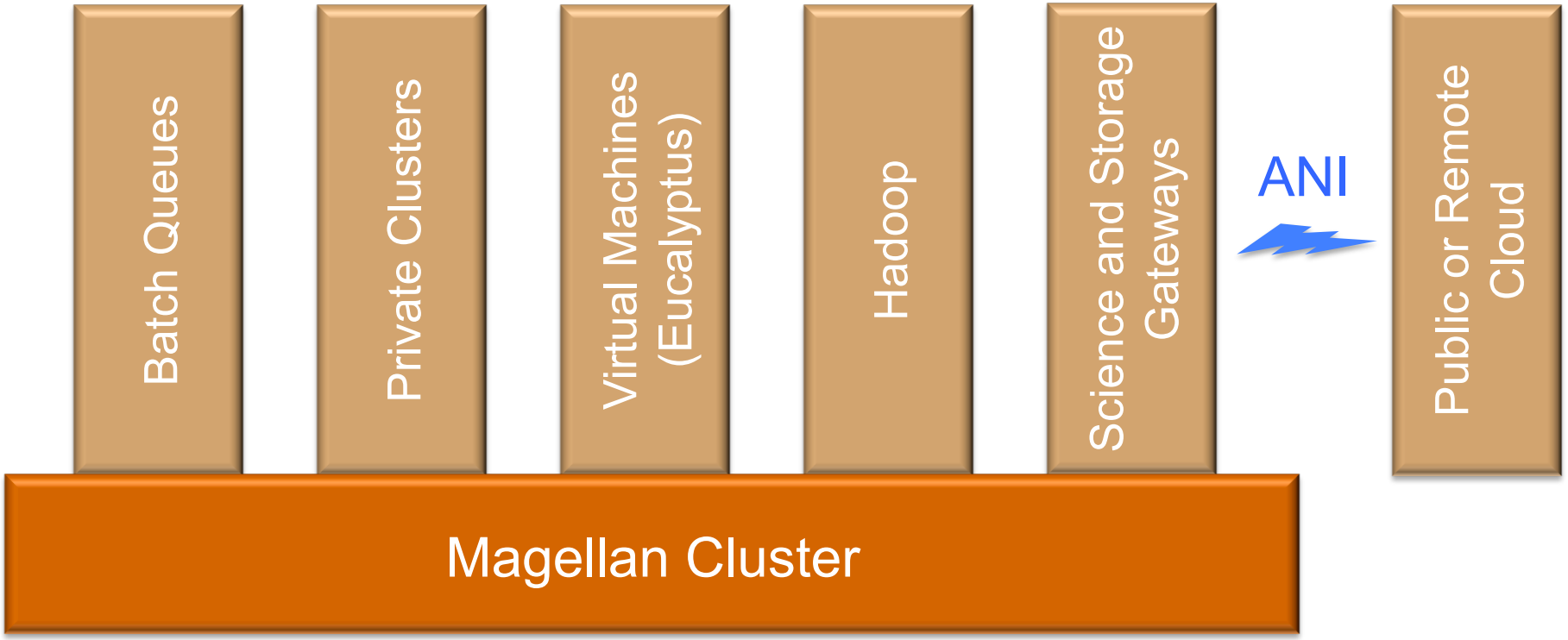
Input Data	Output Data
EBS via NFS	Local storage to EBS via NFS
Staged to local storage from EBS	Local storage to EBS via NFS
EBS via NFS	EBS via NFS
Staged to local storage from EBS	EBS via NFS
EBS via NFS	Local storage to S3
Staged to local storage from S3	Local storage to S3



Total Cost (Instance + Storage/ month)

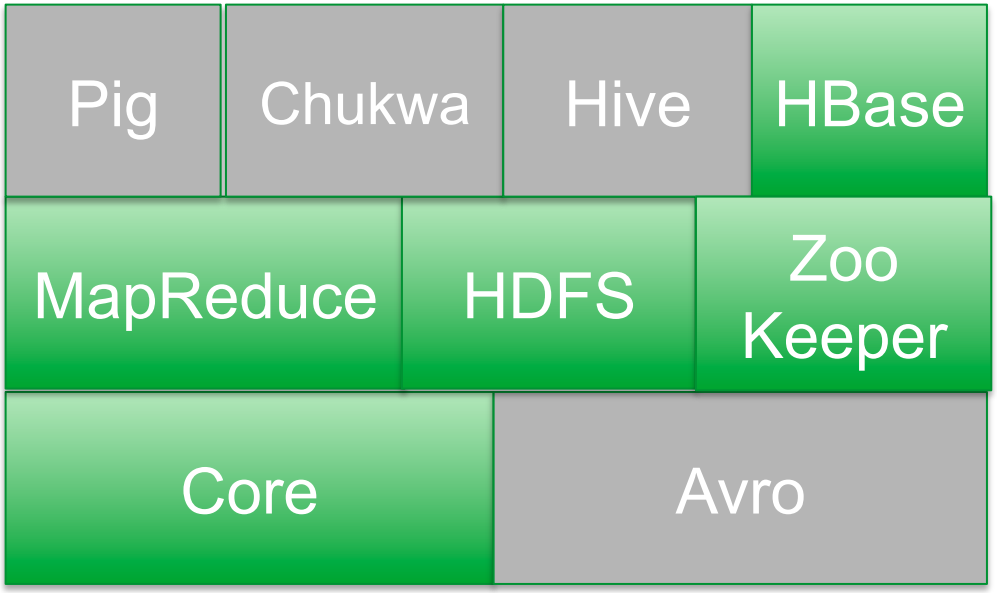


Magellan Software



Hadoop Stack

- **Open source reliable, scalable distributed computing**
  - implementation of MapReduce
  - Hadoop Distributed File System (HDFS)



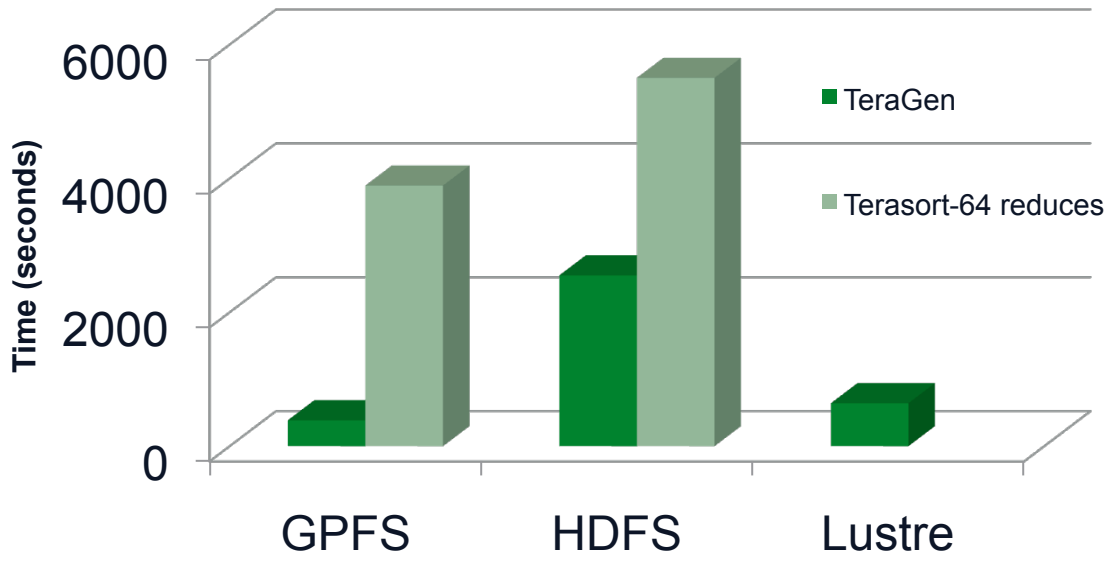
Source: Hadoop: The Definitive Guide

Hadoop for Science

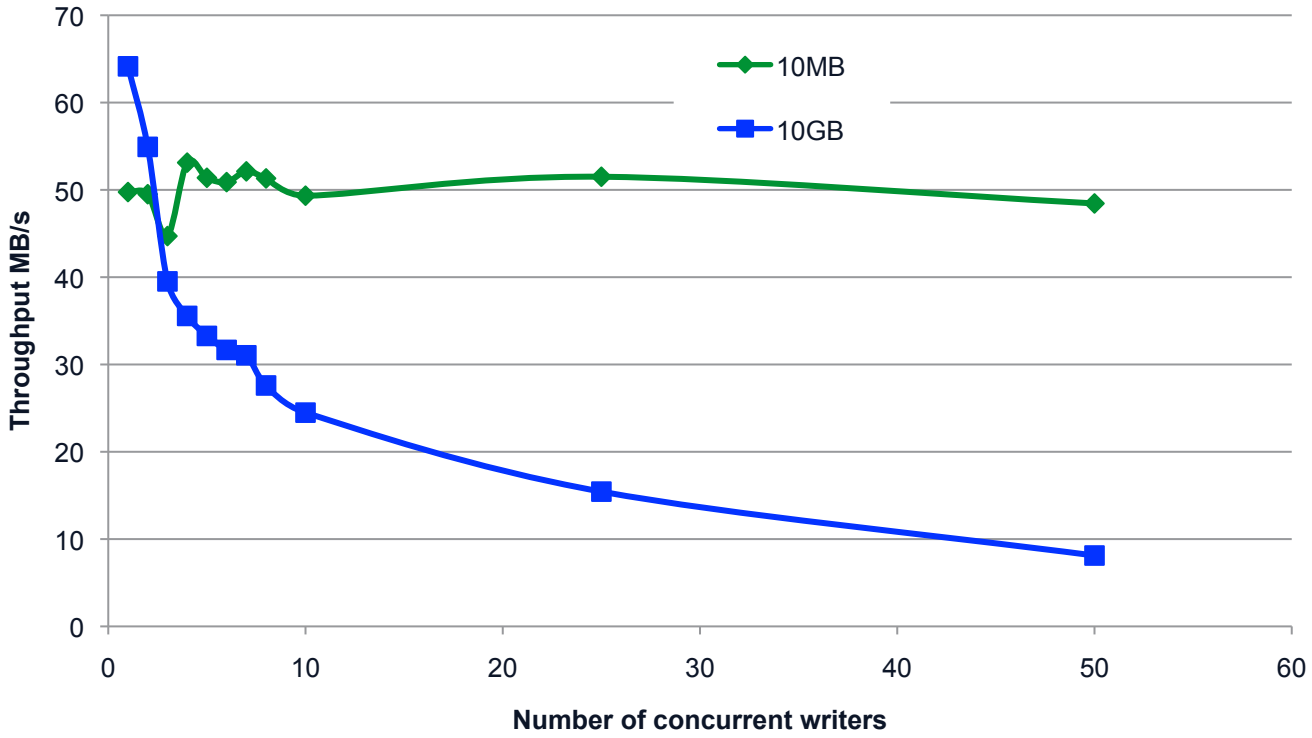
- **Advantages of Hadoop**
  - transparent data replication, data locality aware scheduling
  - fault tolerance capabilities
- **Mode of operation**
  - use streaming to launch a script that calls executable
  - HDFS for input, need shared file system for binary and database
  - input format
    - **handle multi-line inputs (BLAST sequences), binary data (High Energy Physics)**

Hadoop Benchmarking: Early Results [1/2]

- **Compare traditional parallel file systems to HDFS**
  - TeraGen and Terasort to compare file system performance
    - 32 maps for TeraGen and 64 reduces for Terasort over a terabyte of data

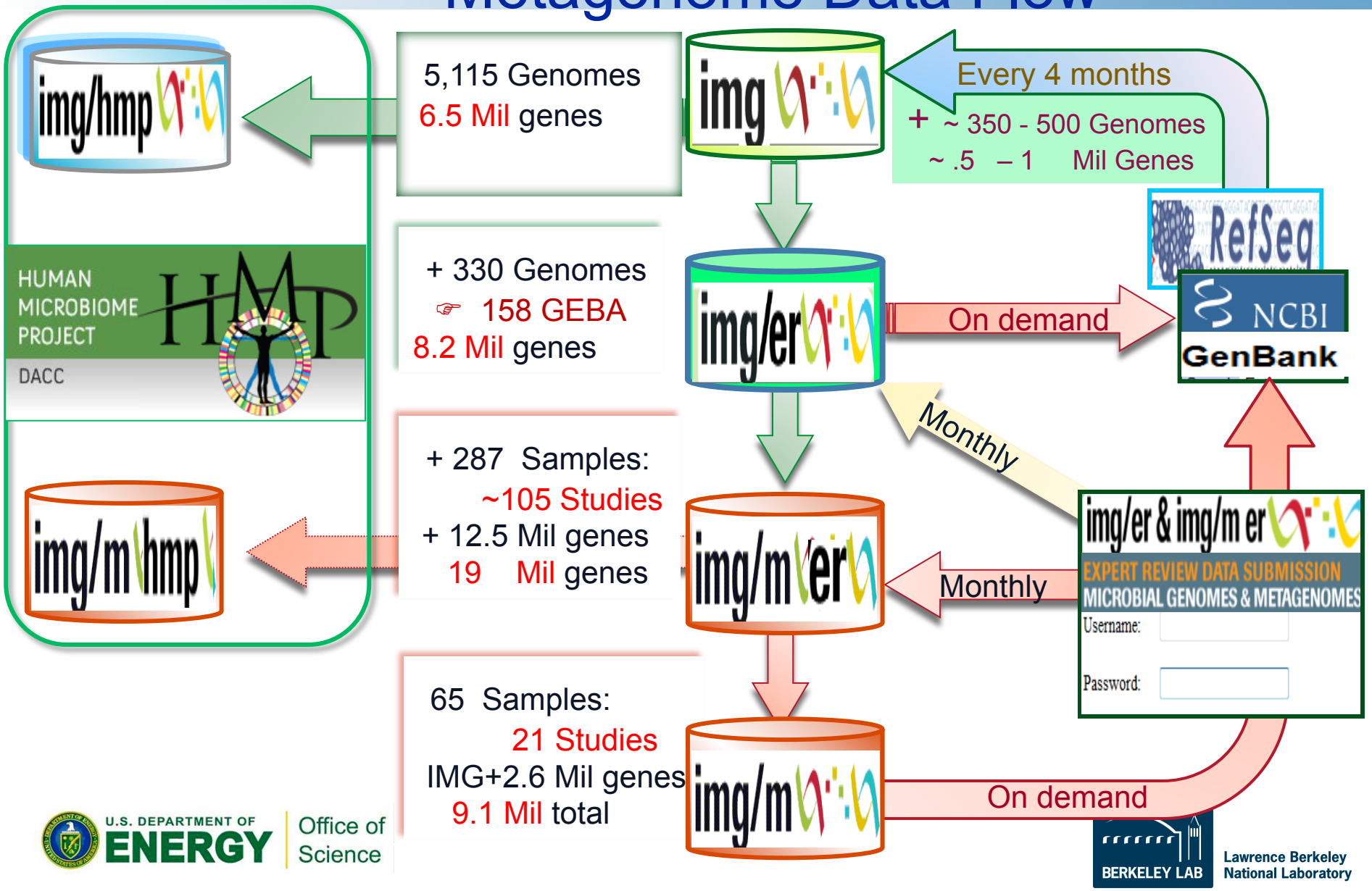


Hadoop Benchmarking: Early Results [2/2]



TestDFSIO to understand concurrency at default block size

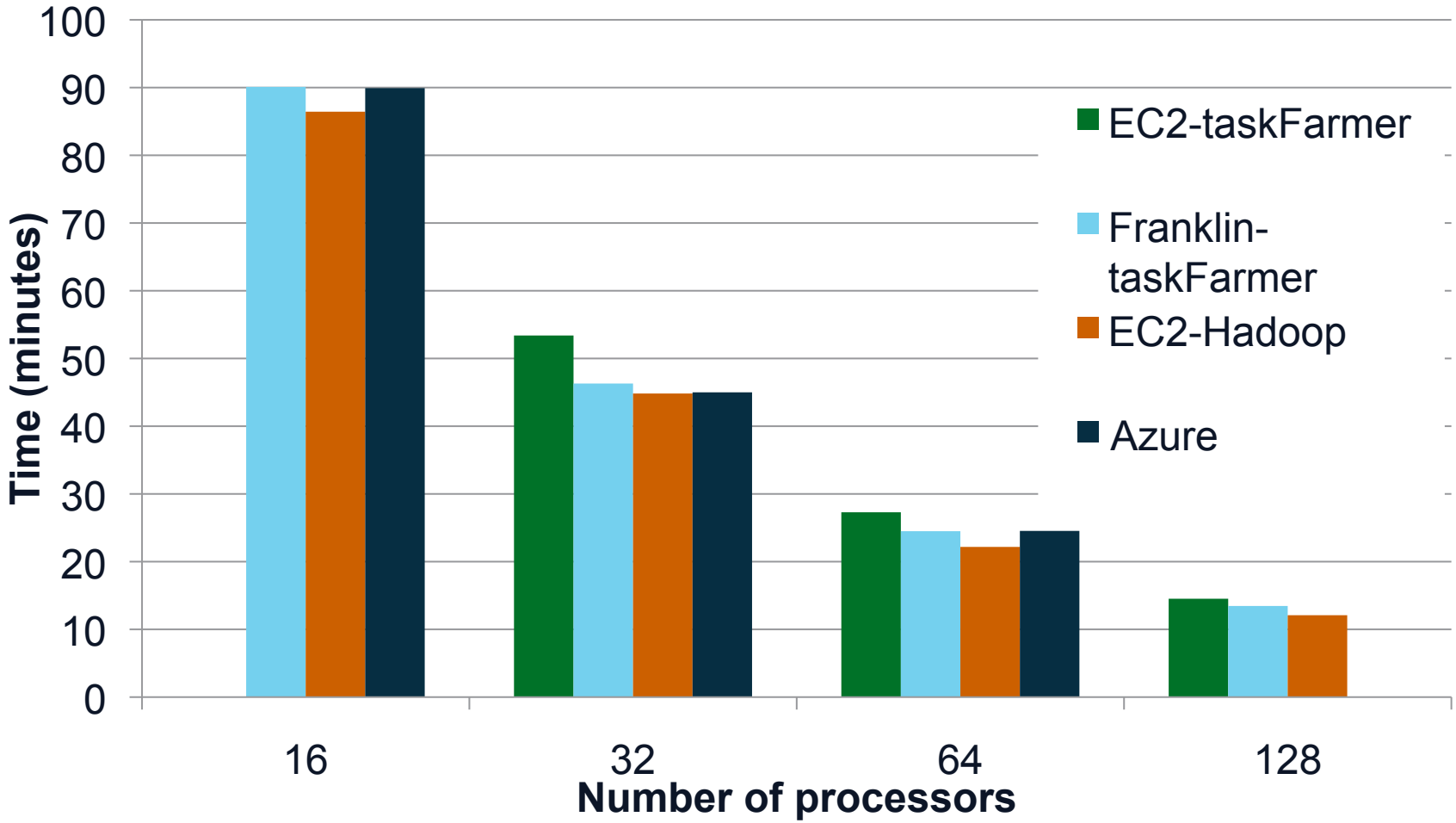
IMG Systems: Genome & Metagenome Data Flow



BLAST on Hadoop

- **NCBI BLAST (2.2.22)**
  - reference IMG genomes- of 6.5 mil genes (~3Gb in size)
  - full input set 12.5 mil metagenome genes against reference
- **BLAST Hadoop**
  - uses streaming to manage input data sequences
  - binary and databases on a shared file system
- **BLAST Task Farming Implementation**
  - server reads inputs and manages the tasks
  - client runs blast, copies database to local disk or ramdisk once on startup, pushes back results
  - advantages: fault-resilient and allows incremental expansion as resources come available

BLAST Performance



BLAST on Yahoo! M45 Hadoop

- **Initial config – Hadoop memory ulimit issues,**
  - Hadoop memory limits increased to accommodate high memory tasks
  - 1 map per node for high memory tasks to reduce contention
  - thrashing when DB does not fit in memory
- **NFS shared file system for common DB**
  - move DB to local nodes (copy to local /tmp).
  - initial copy takes 2 hours, but now BLAST job completes in < 10 minutes
  - performance is equivalent to other cloud environments.
  - future: Experiment with Distributed Cache
- **Time to solution varies - no guarantee of simultaneous availability of resources**

Strong user group and sysadmin support was key in working through this.

Summary: Virtualization for Science

- **Porting applications still requires lots of work**
- **Public clouds**
  - Virtualization has a performance impact
  - Failures when creating large instances
  - Data Costs tend to be overwhelming
- **Eucalyptus**
  - Learning curve and stability at scale
    - Alternate stacks and trying version 2.0
  - SSE instructions were not exposed in VM
- **Additional benchmarking**

Summary: Hadoop for Science

- **Deployment Challenges**
  - all jobs run as user “hadoop” affecting file permissions
  - less control on how many nodes are used - affects allocation policies
  - file system performance for large file sizes
- **Programming Challenges: No turn-key solution**
  - using existing code bases, managing input formats and data
- **Performance**
  - BLAST over Hadoop: performance is comparable to existing systems
  - existing parallel file systems can be used through Hadoop On Demand
- **Additional benchmarking, tuning needed, Plug-ins for Science**

Acknowledgements

This work was funded in part by the Advanced Scientific Computing Research (ASCR) in the DOE Office of Science under contract number DE-C02-05CH11231.

CITRIS/UC, Yahoo M45!, Amazon EC2 Education and Research Grants, Microsoft Research, Wei Lu, Dennis Gannon, Masoud Nikravesh and Greg Bell.

Magellan - Shane Canon, Iwona Sakrejda

Magellan Benchmarking – Shane Canon, Nick Wright

EC2 Benchmarking – Keith Jackson, Krishna Muriki, John Shalf, Shane Canon, Harvey Wasserman, Shreyas Cholia, Nick Wright

BLAST – Shreyas Cholia, Keith Jackson, Shane Canon, John Shalf

Supernova Factory – Keith Jackson, Rollin Thomas, Karl Runge

Questions
LRamakrishnan@lbl.gov

U.S. DEPARTMENT OF ENERGY | Office of Science

Lawrence Berkeley National Laboratory